Reporting uncertainty

Con metodo

Methods

Dolores Catelan, Annibale Biggeri, Fabio Barbone

L'epidemiologia è una delle discipline biomediche che più ha applicato metodi quantitivi. Questa rubrica offre al lettore contributi sugli argomenti e sulle tecniche fondamentali usate attualmente nella produzione scientifica di taglio epidemiologico. Ha pertanto sia uno scopo didattico sia uno scopo innovativo. Infatti alcuni temi classici, fondativi dell'inferenza statistica ed epidemiologica, vengono rivisitati alla luce del buono o cattivo uso che ne è stato fatto. La rubrica non si rivolge solo a chi è più impegnato nella ricerca epidemiologica, ma ha l'obiettivo di fornire strumenti critici anche a un più ampio pubblico di lettori. Particolare attenzione sarà posta all'uso di misure di effetto e di incertezza che abbiano una maggiore valenza comunicativa e permettano meno fraintendimenti e oscurità di interpretazione. Un metodo aperto e in evoluzione, in rapporto con una molteplicità di soggetti e non chiuso nella specificità tecnico-professionale di un singolo ambito disciplinare.

Informiamo i lettori che sul prossimo numero di E&P verrà pubblicata una versione in italiano di questo articolo rivolta a un pubblico non specialistico

Reporting Uncertainy

Annibale Biggeri,[1,2] Dolores Catelan,[1,2] Fabio Barbone[3]

[1] *Department of Statistics "G. Parenti", University of Florence, Florence, Italy*

[2] *Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Florence, Italy*

[3] *Chair of Hygiene and Epidemiology, DPMSC, University of Udine, Udine, Italy*

We are approaching the 25th anniversary of the publication of Martin Gardner's and Douglas Altman's paper on the use of confidence intervals in reporting study results in medical research.[1] Two years after its publication, the International Committee of Medical Journal Editors endorsed this policy.[2,3] Notwithstanding this, there is still a misuse of confidence intervals as a surrogate test of hypothesis and the rhetoric of uncertainty hides uncritical faith on p-values.

In this contribution, we will discuss three issues:

- how to interpret a confidence interval;
- how to report confidence intervals in a paper;
- how to report a confidence interval in an abstract.

How to Interpret a Confidence Interval

A confidence interval is a range of values for a population parameter calculated from a given sample of observations. It is meaningful when we want to make an inference outside our study, which is almost the case in every scientific investigation. In statistical textbooks, the confidence interval is described as an interval estimate of a population parameter. The width of the interval depends on the natural variability of the phenomenon under study, the sample size and the arbitrary level of confidence. Fixing this last quantity, the confidence interval will show the reliability of an estimate and, in a broad sense, the half-width of the confidence interval is called the margin of error. The results of a study, when reported with confidence intervals, have the same standard unit or magnitude of the investigated phenomena. This was the main justification evoked for a shift from reporting standardized effect measures or their statistical significance. The advantage of using a confidence interval is that the degree of uncertainty is translated into the width of the interval and anyone can immediately appraise how informative a study result is and determine the weakness related to a small sample size or lack of control of population variability in the factor being studied. The "interpretation of confidence intervals should

focus on the implications (clinical importance) of the range of values in the interval".[4]

There is an inherent arbitrariness in the specification of the level of confidence used to calculate the confidence interval. Under some assumptions, there is a correspondence between the test of hypothesis and the interval estimate. Therefore, if we claim results statistically significant at 5% two-sided, then the 95% confidence interval will exclude the null value. This is true provided that the assumptions be satisfied, but the unwanted consequence is that too often in biomedical research confidence intervals are judged only on the basis of the criterion of excluding the null value. Warning against the acritical use of any pre-fixed level of confidence was expressed.[5] In order to discourage improper interpretation of confidence intervals, Sterne and Davey-Smith[4] suggested reporting intervals at the 90% confidence level.

The frequentist derivation of the confidence interval assumes a priori infinite repetitions of the study with the same fixed sample size. For each replicate we calculate a confidence interval and by random sampling we select just one of them. Under a Gaussian probability model, we can build confidence intervals with a width that provides a given probability of selecting an interval that includes the population parameter. Under such a paradigm, once having done the study and having estimated one confidence interval, any value in it has the same probability to be equal to the population parameter. This reflects our ignorance and the exchangeability of the replicates under the random sampling paradigm. The Gaussian probability model also gives an interpretation to the arbitrary cut-off used for the confidence level: selecting 95% confidence level is the equivalent of saying that one over twenty sampled confidence intervals will exclude the population parameter.

Using likelihood theory, an interval estimate for a population parameter or *supported range* is the set of values of the parameter with likelihood ratios above a critical value. Having a probability model, e.g. a Bernoulli model for binary data or a Poisson model for disease event counts, we can calculate the data likelihood and derive an appropriate supported range from the profile likelihood ratio function of the parameter of interest.[6]

An example

Let us consider the point source study on the high frequency radio transmitter in Rome and the incidence of childhood leukemia (table 4, modified):[7]

distance from source	Incident cases of childhood leukemia
0-2 km	1 observed vs 0.16 expected
0-6 km	8 observed vs 3.68 expected

Table 1. Childhood leukemia incident cases and expected counts by distance from putative source (see text).

The probability model is Poisson: the disease counts Y are assumed to be distributed as a Poisson random variable with the mean parameter equal to $\theta \times E$; E being the expected count under indirect standardization:

$$Pr(Y = y|\theta\, E) = C\theta^{Y} exp(-\theta E)$$

where C is a constant ($E^{Y}/Y!$). The log likelihood function reduces to:

$$l(\theta) = Y log(\theta) - \theta E$$

For the first row in the table above, the standardized incidence ratio is 1/0.16=6.25 and it is the maximum likelihood estimate of the relative risk θ. In fact it is the value which corresponds to the maximum of the likelihood function
(1 x log (6.25)-0.16 x (6.25) = 0.83258146).
A supported range is calculated by finding the values which satisfy:

$$l(\theta) = Y log(\theta) - \theta E = 1.353$$

where the cut-off value of -1.353 is arbitrary.
We found that the equation is satisfied for θ = 0.6605; 22.7930.
The log likelihood ratio (i.e. the log likelihood minus the maximum of the function) is indeed:

(1 x log (22.7930)-0.16 x (22.7930) - 0.83258146 = -1.353
(1 x log (0.6605)-0.16 x (0.6605) - 0.83258146 = -1.353

The same calculations for the second row in the table will give θ = 1.14185; 3.69435.
While for the first row the empirical evidence was inconclusive, because the supported range was very wide with non-sensible values from 0.66 to 22.79 , the data for the 0-6km band supported relative risk in the range 1.14 ÷ 3.69, a range of values consistent with epidemiological literature on environmental exposures. However, the causal interpretation of such findings is not a statistical issue.
In this example we used the cut-off of -1.353 for the log likelihood ratio. Under a Gaussian approximation to the likelihood and by applying the frequentist approach, it would correspond to a 90% confidence level. The relationship is -2 log likelihood ratio = z^2 and we obtain for example:

(-2) x (-1.353)=1.645^2 for a 90% confidence level
(-2) x (-1.921)=1.96^2 for a 95% confidence level.

The usual approximate formula for the confidence interval is:

$$ss \pm z_{1-\alpha/2} \times se(ss)$$

where ss is the generic sample statistics, for example the sample average, $z_{1-\alpha/2}$ is an appropriate centile of a theoretical sampling distribution, for example the normal or the student's t,

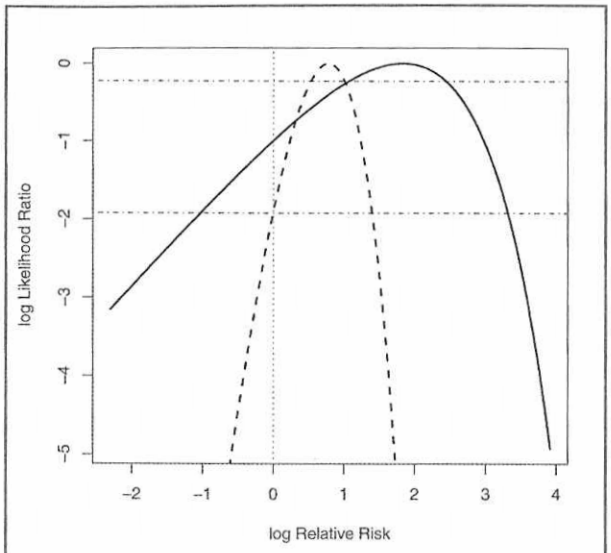

Figure 1. log Likelihood ratio function for leukemia example (see text). Solid line: 1 case, 0.16 expected; dashed line: 8 cases, 3.68 expected; dotdash lines: cut-off values at -0.23 and -1.92 corresponding to standard normal confidence levels of 50% and 95%.

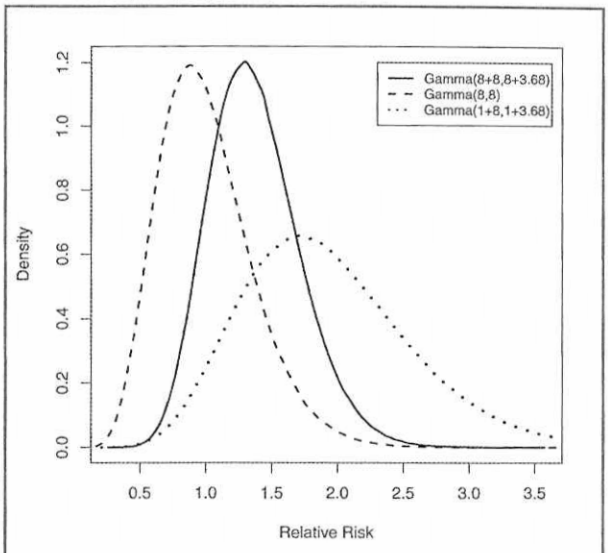

Figure 2. Gamma(8,8) prior, posterior Gamma(8+8,8+3.68) and the posterior Gamma(1+8,1+3.68).

and *se* is the standard error of *ss*. In the case of the standardized ratio example above, the Gaussian approximation would be $ss = Y/E$ and $se(ss) = \sqrt{Y}/E$ and we obtain

$1/0.16 \pm 1.645 \times 1/0.16 = -4.03; 16.53$ and
$8/3.68 \pm 1.645 \times \sqrt{8}/3.68 = -0.91; 3.44$.

The Poisson likelihood for the small observed number of cases is strongly asymmetric and the approximation is not adequate. In the biostatistical literature, several approaches to confidence interval estimation are discussed.[8] We aimed here only to reinforce the message that the empirical evidence supports a *range* of plausible values for the parameter (effect measure) of interest. The uncertainty in empirical research implies that we should scrutinize not one solution but a portfolio of alternatives.

How to Report a Confidence Interval in a Paper

It is common practice in epidemiological literature to report confidence intervals (CIs) after point estimates, as for example (90% CI: low ; up). This may be confusing in two ways as shown in Louis and Zeger.[9] In the tables it could be difficult to directly compare point estimates, because of confidence intervals in adjacent columns. This seems to be a minor problem, but it underlines the fact that the researcher is pushed to consider separately point and interval estimates. Instead, both of them are summaries of the same information, which is driven by the data likelihood function. In Figure 1, we report the two log likelihood ratio functions for data reported in Table 1. The curves show that the larger the sample size (8 cases *vs* 1 case) the more peaked the likelihood and the shorter the supported range. Moreover, they illustrate the second argument of Louis and Zeger:[9] the likelihood is not the same for all the points (relative risks) in the supported range. The authors proposed to report the maximum likelihood estimate together with supported ranges corresponding to a confidence level of 50% and 95%, i.e. using the 25%-75% and 2.5%-97.5% centiles of the standard normal. Using again data from Table 1, the way to report the point estimate and the related uncertainty should be for the first relative risk (1 case event *vs* 0.16 expected) $_{0.36}\ _{2.92}\ 6.25\ _{11.47}\ _{27.66}$ and for the second (8 case events *vs* 3.68 expected) $_{0.99}\ _{1.70}\ 2.18\ _{2.73}\ _{4.06}$. A simpler solution would be to report only one supported range corresponding to confidence levels of 90% or 95%, e.g. for a 90% supported range we get $_{0.66}\ 6.25\ _{22.79}$ and $_{1.14}\ 2.18\ _{3.69}$ respectively.

The maximum likelihood estimate is written in full text and the limits of the supported range as left and right sub-indices, recursively if more than one supported range is reported. One great advantage of this solution, if it is sensible for the problem, is that the reader has an idea of the location of the whole likelihood function with respect to the null relative risk value. In the first case (1 case event *vs* 0.16 expected) the empirical evidence quantified by the likelihood ratio function is largely concentrated on relative risks above the RR=1, information completely lost when reporting the two confidence limits alone – see also Rothman for a discussion of this problem.[10]

Box 1. Bayesian inference is based on the posterior distribution:

$$\Pr(\theta \mid Y) = \frac{\Pr(Y \mid \theta)\Pr(\theta)}{\Pr(Y)}$$

which is derived by applying the Bayes formula. $\Pr(Y \mid \theta)$ is the likelihood and $\Pr(\theta)$ is the prior.
This formula is used also for density functions. In the example shown in the text we have:

$$\Pr(\theta \mid Y) = \frac{\Pr(Y \mid \theta)\Pr(\theta)}{\Pr(Y)} = \frac{Poisson\,(Y;\, E,\theta)\; x\; Gamma\,(\theta;\, a,\, b)}{C} = Gamma\,(\theta;\, a + Y, b + E)$$

This approach suffers from both the arbitrariness in the definition of the confidence levels and the Gaussian assumptions for their interpretation according to the frequentist paradigm. A Bayesian approach will provide a credibility interval which is simpler and easier to understand.[11] Combining the data likelihood with the prior distribution, Bayesian inference uses the posterior distribution for the parameter of interest (see box 1). This is a probability distribution and it is summarized reporting a measure of central tendency (i.e. the mode, the mean or the median) together with selected centiles of the distribution. This is called the credibility interval and the associated level is the probability that the parameter of interest, given the data, has a value in that interval. Let's consider the leukemia data of Table 1: 8 case events *vs* 3.68 expected and the supported ranges (50% 95%): $_{0.99}\ _{1.70}\ 2.18\ _{2.73}\ _{4.06}$. Using, for mathematical convenience, the Gamma prior, which is conjugate to the Poisson distribution, we can derive in a closed form the posterior distribution: for example if the prior is a Gamma(a,b) the posterior still will be a Gamma with parameter (a+Y,b+E), Y and E being the observed number of event cases and the expected counts, respectively. In Figure 2, we show a Gamma(8,8) prior (dashed black), the posterior Gamma(8+8,8+3.68) and the posterior Gamma(1+8,1+3.68) assuming a much more dispersed prior Gamma(1,1).
The centiles (2.5% 25% mean 75% 97.5%) of these posterior distributions are $_{0.79}\ _{1.14}\ 1.38\ _{1.59}\ _{2.19}$ and $_{0.87}\ _{1.46}\ 1.95\ _{2.38}\ _{3.42}$.

These credibility intervals are directly interpretable: there is a 95% probability that the true unknown relative risk lies in the interval 0.87 : 3.42 under the more dispersed Gamma(1,1) prior, and there is a 90% probability that it lies in the interval 0.99 : 3.15 . This example is useful also to underline the role of prior belief. If we choose the Gamma(8,8) prior, we are assuming that the range of plausible relative risks would be (2.5% 25% mean 75% 97.5%) $_{0.42}\ _{0.73}\ 1.00\ _{1.18}\ _{1.80}$ (see[6] pages 117-119 for a discussion about the choice of prior distributions). In such a case, given the small number of events, the posterior is shifted more toward the null relative risk of one.[12]

How to Report a Confidence Interval in an Abstract

Recently, there was a debate about the potential pitfalls of epidemiologic research. The credibility of scientific investigations was questioned daily by the reports in the media of unconfirmed new risk factors.[13] Sterne and Davey-Smith issued warnings against subgroup analysis to limit discredit to epidemiological research.[4] In fact, current epidemiological studies have large sample sizes and focus on small risk factors for population subgroups. These studies usually assess several research hypotheses, even thousands or millions in Genome-wide analysis. The papers contain large tables of relative risks and confidence intervals. (see[14] for examples in descriptive epidemiology) Leaving aside the problem of testing multiple hypotheses, which is typical in Genomics, here we address the arbitrariness in reporting only some results in an abstract. This is important for two reasons: abstracts are open access and may influence a large audience; there is room for arbitrary selection of the research findings.
The coverage of the confidence interval under selection has been shown to be invalid. When we select from an abstract some relative risks and their confidence intervals from a large set reported in the body of the text, the width of those confidence intervals is too short and need to be adjusted for the selection process.[15] Suppose we aim to study population susceptibility to air pollutants and provide a list of 20 relative risks (and confidence intervals) in the body of the manuscript. We then select the two more important relative risks (and confidence intervals) to appear in the abstract. Now, while the confidence intervals given in the body of the text are valid, since they are listed together with all the others, the same cannot be said of the two reported in the abstract. The uncertainty due to the selection process is not accounted for. The correction suggested by Benjamini and Yekutieli[15] is simple:

$$ss \pm z_{1-\alpha'/2} \times se$$

$$\alpha' = R \times \frac{\alpha}{m}$$

α is the desired confidence level, R is the number of selected confidence intervals to be reported in the abstract and m is the

total number of confidence intervals in the manuscript. Then, if we report two confidence intervals out of twenty the centile of the sampling distribution should be chosen for 2 x $\alpha/20$ – i.e. for a 90% confidence interval we must use $z_{1-\alpha'/2} = 2.576$ instead of 1.645. Let us explain in detail.

Suppose we calculate 100 confidence intervals at confidence level a and choose to report in the abstract those intervals that exclude the null value. Then the conditional coverage probability – $Pr(\theta \in CI \mid CI$ selected), which the number of times a confidence interval includes the parameter divided by the number of times the confidence interval is selected, is no longer fixed to α. As shown in Benjamini and Yekutieli,[15] it varies and depends on the value of the unknown parameter being estimated. Defining the False Coverage Rate as the number of times a confidence interval *does not* include the parameter divided by the number of times the confidence interval is selected we can provide a way to properly control its expected value, having set to zero the proportion when no CI is selected. For example, suppose that the true value of the relative risk parameter be one, and *m* confidence intervals be calculated. Then select a confidence interval if it excludes the null value RR=1. The conditional coverage probability is zero and the expected False Coverage Rate is one. However, if we modify our selection procedure using for example the Bonferroni correction, i.e. setting the confidence level at $\alpha'=\alpha/m$, the expected False Coverage Rate is α. Formally:

$$E\left(\frac{V}{R} \mid R > 0\right) Pr\,(R > 0)$$

V are the number of false coverages, *R* the number of selected CI. In the situation described before (RR=1), averaging over repetitions, the first factor will be one, as before, while the second factor will be exactly 0.05 , the family-wise error rate assured by the Bonferroni correction. This procedure is too strict whenever, for some relative risks, the null hypothesis RR=1 is false. This justifies the formula reported above, which substitutes $\alpha' = \alpha/m$ with $\alpha' = R \times \frac{\alpha}{m}$

Conclusion

In this contribution, we discussed three issues: the interpretation to be given to a confidence interval; a proposal to report confidence intervals in a paper; a suggestion for reporting a confidence interval in an abstract.

The interpretation may be unfamiliar and stresses the importance of the supported range and tries to weaken the connection between confidence intervals and the test of hypothesis.

The proposal may seem awkward to implement in writing a paper, but it emphasizes the importance of a continuum between point and interval estimates and introduces the concept of a distribution on the parameter of interest. The Bayesian approach is natural from this point of view.

The suggestion may seem provocative. We think that it is very important to limit the amount of data dredging in epidemiological research while preserving its power.

We hope our suggestion stimulates the debate and avoids uncritical use of statistics in scientific literature.

Bibliografia/References

1 Gardner MJ, Altman DG. Confidence Intervals Rather Than P values: Estimation Rather Than Hypothesis Testing. *Br Med J (Clin Res Ed)* 1986; 292(6522): 746-50.

2. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *Br Med J* 1988; 296: 401-5.
3. Gardner MJ, Altman DG. Estimating with Confidence. *Br Med J (Clin Res Ed)* 1988; 296(6631): 1210-11.
4. Stern JAC, Smith DG. Sifting the Evidence. What's Wrong with Significance Tests? *BMJ* 2001; 322: 226-31.
5. Gardner MJ, Altman DG. Using Confidence Intervals. *Lancet* 1987; 1(8535): 746.
6. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford, Oxford University Press, 1993.
7. Michelozzi P, Capon A, Kirchmayer U et al. Mortality from Leukemia and Incidence of Childhood Leukemia Near a High Power Radio Station in Rome, Italy. *American Journal of Epidemiology* 2002; 155(12): 1096-103.
8. van Belle G, Fisher L, Heagerty PJ, Lumley T. *Biostatistics: A Methodology for the Health Sciences* (2nd Edition). New York, Wiley, 2004.
9. Louis TA, Zeger SL. Effective Communication of Standard Errors and Confidence Intervals. *Biostatistics* 2009; 10(1): 1-2.
10. Rothman KJ. *Epidemiology: An Introduction.* Oxford, Oxford University Press, 2002.
11. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. (2nd Edition) Chapman & Hall/CRC Press, Boca Raton: 2003
12. Clayton DG, Kaldor J, 1987. Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. *Biometrics* 1987; 43: 671-81.
13. Traubes G. Epidemiology faces its limits. *Science* 1995; 269: 164-69.
14. Catelan D, Biggeri A. Multiple Testing in Descriptive Epidemiology. *GeoSpatial Health* 2010; 4(2): 219-29.
15. Benjamini Y, Yekutieli D. False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters. *JASA* 2005; 100(469): 71-81.